

Doubly-nonparametric generalized linear models

Alan Huang
University of Queensland

4 Dec, 2014

Abstract

We extend nonparametric generalized linear models to allow both the mean curve and the response distribution to be nonparametric. The seemingly intractable task of working with two infinite-dimensional parameters is shown to be reducible to a finite optimization problem, which is easily implemented via existing algorithms. We demonstrate using various examples that the proposed approach can be a flexible tool for data analysis in its own right, but can also be useful for model selection and diagnosis in a more classical generalized linear model framework.

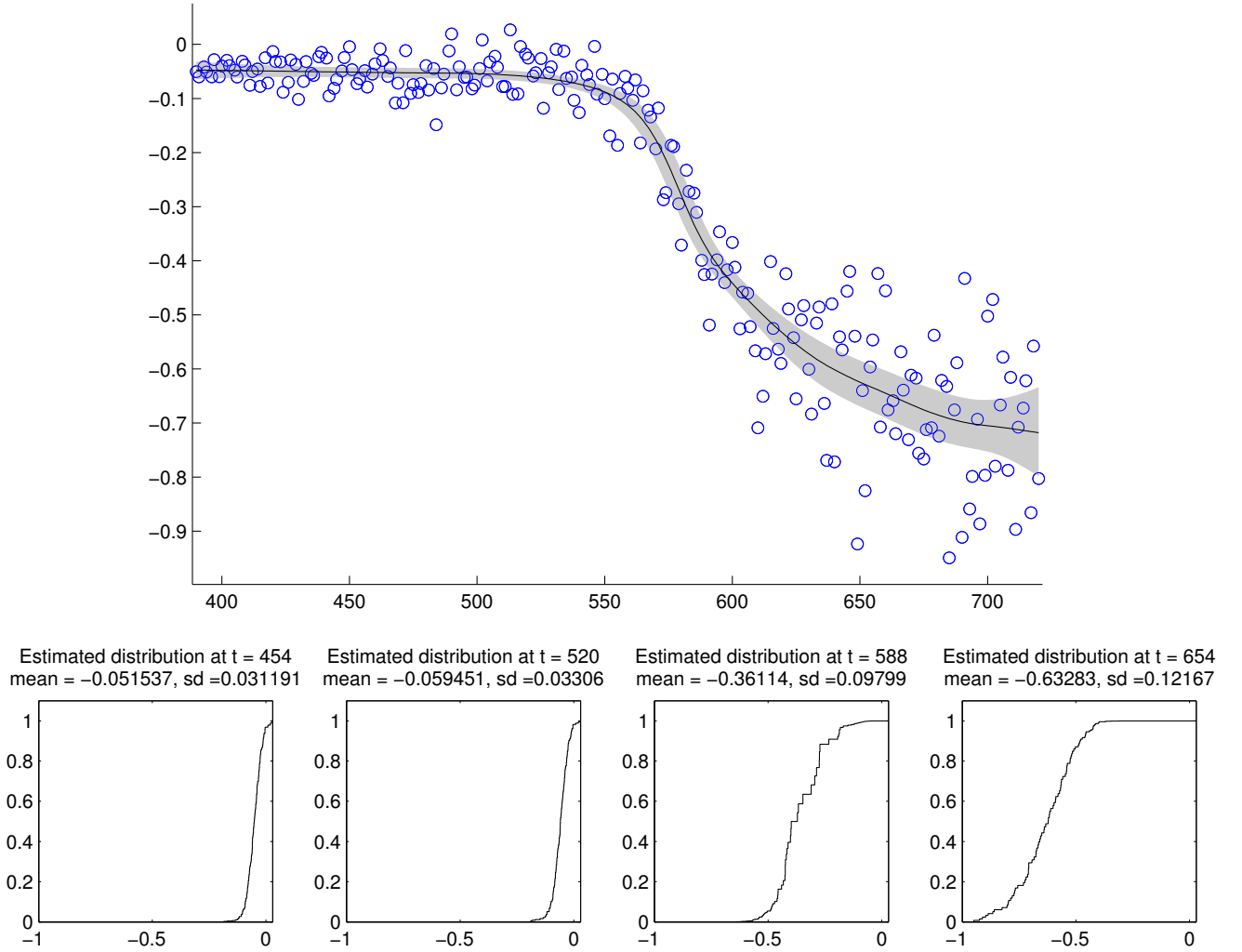
Keywords: Generalized linear models; Scatterplot smoothing; Smoothing splines.

1 Introduction

Nonparametric generalized linear models (e.g., [Green & Silverman, 1994](#), Chapter 5; [Ruppert, Wand & Carroll 2003](#), Chapter 10) have become popular tools for applied regression analysis because of their ability to model a wide range of data and patterns in a flexible, nonparametric way. The phrase “nonparametric generalized linear model”, is, however, somewhat of a misnomer because such models are really only semiparametric in nature. This is because while the mean curve is nonparametric, the error distribution around the mean (or, equivalently, the variance function) is typically fully specified, except perhaps for a finite number of scale parameters. It seems rather paradoxical to be entertaining such flexible curves for the mean yet remain so rigid with the response distribution. Indeed, it is well-known that even in simple linear regression settings, misspecification of the response distribution can lead to significantly biased inferences (e.g. [Eicker, 1967](#); [White, 1982](#)). Our simulation results in Section 6 suggest that model misspecification can be equally, if not more, detrimental in nonparametric regression settings. The problem is further amplified in small to moderate sized problems where accurate model selection and diagnosis can be difficult precisely because of the lack of information.

This paper introduces a novel extension of nonparametric generalized linear models (GLMs) that allows both the mean curve and the response distribution to be nonparametric. The proposed approach is a genuine extension of GLMs, in that the only model assumption we make is that the data come from *some* exponential family. Consequently, we always remain in a full probability setting. In contrast, quasilikelihood (QL) based methods, such as QL with unknown link and variance ([Chiou & Müller, 1998](#)), typically do not correspond to any probability model for the data and thus do not provide any further insight into the probabilistic mechanism generating the data beyond that of the first two moments. Having a full probability model is particularly useful for model selection and diagnosis, predictive inferences and nonparametric bootstrap resampling.

Figure 1: Main display: Scatterplot of the original data (circles) with estimated mean curve (solid) and approximate pointwise 95% variability band for the mean (shaded) for the LIDAR dataset. Second row: Estimated response distributions at quintiles of the covariate values, along with the estimated means and standard deviations. Sample size $n = 221$, smoothing parameter obtained by two-fold cross-validation



In Figure 1, we display the output from a MATLAB routine that fits the proposed model to a light detection and ranging (LIDAR) dataset from [Ruppert, Wand & Carroll \(2003\)](#). The main features of this dataset are the nonlinear relationship between the response and the covariate and the heteroscedastic errors, making it difficult to model using classical regression approaches. Here, the response (y) is the logarithm of the ratio of received light from two laser sources and the covariate (t) is the distance traveled before the light is reflected back to its source.

In the main display of Figure 1, the circles represent the data points, the solid line is the fitted mean curve and the shaded area is an approximate 95% pointwise variability band for the mean. We see that the nonlinear relationship between the response and covariate has been captured well by the fitted nonparametric mean curve, with the variability band suitably wider in regions where the variance is evidently larger. In the second row of the figure, the estimated response distributions at quintiles of the covariate values are displayed, along with their corresponding means and standard deviations. We see that both the shape and the spread of the fitted response distributions change as a function of the covariate value, following the non-constant variance patterns apparent in the data. Note that both the mean curve and the response distributions were estimated simultaneously and automatically, and

at no stage did we specify what kind of patterns to look for.

The only aspect of the proposed approach that requires user input is the selection of a smoothing parameter. However, this process can also be automated via a selection method such as cross-validation (see Section 8). Note that smoothing parameters are central to all smoothing methods in statistics, including the classical approach of [Green & Silverman \(1994\)](#) which this paper extends. An attractive aspect of our proposed approach is that it only requires a smoothing parameter, whereas existing methods require specification of both a smoothing parameter and an underlying response distribution for the data. As is evidenced through the introductory example in Figure 1 and the various synthetic examples in Section 6, the relaxing of distributional assumptions makes the doubly-nonparametric GLM approach a powerful and flexible tool for data analysis.

The proposed approach also possesses two very attractive theoretical properties that are generally not true for classical nonparametric GLMs. First, it automatically preserves support bounds, without requiring these bounds to be known *a priori*. Thus, if the data are non-negative, say, then the fitted mean curves and confidence bands are automatically non-negative also. Second, the proposed approach is invariant to shifting and scaling of the data. Thus, fitted mean curves and confidence bands for linearly transformed data are precisely the transformed versions of the fitted means and confidence bands for the original data. Finally, the proposed approach is easily extendible to multiple covariates, thereby extending the generalized additive framework of [Hastie & Tibshirani \(1990\)](#) as well.

2 Model and method

2.1 Classical nonparametric GLMs

We first review the classical penalized likelihood approach to nonparametric GLMs of [Green & Silverman \(1994\)](#). The extension to doubly-nonparametric GLMs is then developed using a novel exponential tilt representation of GLMs introduced in [Rathouz & Gao \(2009\)](#). For simplicity, we focus on the case where there is only one quantitative covariate, denoted generically by t , with the observed covariate values being distinct and ordered, $t_1 < t_2 < \dots < t_n$; extensions of this framework are discussed in Section 7. The responses are denoted generically by Y , with a particular sample is denoted by $\mathbf{Y} = (Y_1, \dots, Y_n)^T$.

Following Chapter 5 of [Green & Silverman \(1994\)](#), recall that a nonparametric GLM assumes that the responses Y_i , conditional on the covariates t_i , are independently drawn from an exponential family of distributions with densities of the form

$$dF_i(y; g(t_i); \phi) = \exp \left\{ \frac{g(t_i)y - b(g(t_i))}{\phi} + c(y, \phi) \right\}, \quad (1)$$

where ϕ is a scale parameter, the functions b and c are known functions that determine the form of the distribution, and g is a smooth but otherwise unspecified function. Note that the mean function is $E(Y|t) = \mu(t) = \int y dF(y; g(t), \phi) = b'(g(t))$, a smooth function of t , related to g via the canonical link b' .

The loglikelihood function for g is given by

$$l(g) = \frac{1}{n} \sum_{i=1}^n \{g(t_i)Y_i - b(g(t_i))\},$$

up to constants not involving g . It is clear that attempting to maximize this loglikelihood function over all smooth functions g is useless, since it is always possible to choose g sufficiently complicated that it interpolates the data, resulting in an arbitrarily large loglikelihood value. An intuitive remedy is to add an additional term to the loglikelihood that penalizes

functions g that are overly complicated. Perhaps the most popular and arguably most elegant way of doing this is that of [Green & Silverman \(1994\)](#), which uses a penalty term proportional to $\int \{g''(t)\}^2 dt$. The relative merits of this penalty term are well-discussed in the literature, for example, in [Green & Silverman \(1994, Section 3.8\)](#) and in [Ruppert, Wand & Carroll \(2003, Section 3.15\)](#). The resulting penalized loglikelihood function is now of the form

$$l_\alpha(g) = \frac{1}{n} \sum_{i=1}^n \{g(t_i)Y_i - b(g(t_i))\} - \frac{1}{2}\alpha \int \{g''(t)\}^2 dt, \quad (2)$$

where $\alpha \geq 0$ is a smoothing parameter. Small values of α lead to mean curves that adhere quite closely to the data, whereas increasing the value of α leads to increasingly linear functions g on the canonical scale. How this translates to the smoothness of the mean curve depends on the link function, b' , which is in turn determined by the assumed distribution.

Maximizing (2) over functions g in the space of all continuous, twice-differentiable functions with absolutely continuously first derivatives leads to a *smoothing spline estimator* of g , also known as the *natural cubic spline estimator*. For more discussion on the theoretical and practical properties of smoothing splines, see [Green & Silverman \(1994\)](#).

2.2 Doubly-nonparametric GLMs

A recent innovation by [Rathouz & Gao \(2009\)](#) showed that any family of distributions with densities of the form (1) can be rewritten as

$$dF_i(y) = \exp \{-b_i + g(t_i)y\} dF(y)$$

for some reference distribution F , where b_i are normalizing constants given by

$$b_i \equiv b(g(t_i), F) = \log \int \exp \{g(t_i)y\} dF(y), \quad i = 1, \dots, n. \quad (3)$$

In other words, each density dF_i is an exponential tilt of the reference density dF , with the amount of tilting determined by the covariate value t_i through $g(t_i)$. Note that the scale parameter ϕ has been absorbed into the functions b, g and F . Note also that, as with any GLM, the distribution F must have a Laplace transform in some neighborhood of 0, so that the cumulant generating function in (3) is well-defined.

The key advantage of this exponential tilt representation is that it makes apparent the idea that the reference distribution F can itself be considered an infinite-dimensional parameter in the model. Indeed, we can now write the loglikelihood as a function of both g and F ,

$$l(g, F) = \frac{1}{n} \sum_{i=1}^n \{\log dF(Y_i) - b(g(t_i), F) + g(t_i)Y_i\}, \quad (4)$$

with $b(g(t_i), F)$ given by (3). Treating F as a free parameter introduces much flexibility and robustness into the model. For example, overdispersed count data can be dealt with simply by F having heavier tails than a Poisson distribution. Similarly, zero-inflated counts can be dealt with simply by F having excess probability mass at zero. Most remarkable, perhaps, is that F can be left completely unspecified and estimated nonparametrically from the data, along with the mean-curve. That is, we let the data inform us as to which mean-curve and response distribution fit best.

As with classical nonparametric GLMs, the loglikelihood function (4) can be made arbitrarily large for any given F by choosing g to be sufficiently complicated. Hence, we consider a doubly-nonparametric analogue of the penalized loglikelihood function (2) given by

$$l_\alpha(g, F | \mathbf{t}, \mathbf{Y}) = \frac{1}{n} \sum_{i=1}^n \{\log dF(Y_i) - b(g(t_i), F) + g(t_i)Y_i\} - \frac{1}{2}\alpha \int \{g''(t)\}^2 dt, \quad (5)$$

with $b(g(t_i), F)$ given by (3).

A doubly-nonparametric estimator of g and F can then be defined as the joint maximizer of the penalized loglikelihood (5), where the maximization in g is taken over \mathcal{S}_2 , the space of all continuous, twice-differentiable functions with absolutely continuously first derivatives, and the maximization in F is taken over \mathcal{F} , the space of all distributions having a Laplace transform in some neighborhood of 0. Working in this doubly infinite-dimensional space appears at first to be an intractable task, but as we show in Proposition 1 below, this is reducible to a finite maximization problem involving at most $2n - 1$ variables, where n is the sample size. As a comparison, simultaneous estimation of the mean and the variance function using smoothing splines generally requires an optimization over $2n$ variables, which is of the same order. Note that all functions involved in the simplified optimization problem of Proposition 1 below have explicit forms, making it easily implementable in mathematical computing software such as MATLAB.

Proposition 1. Denote by (\hat{g}, \hat{F}) the maximizer of the penalized likelihood (5) over $\mathcal{S}_2 \times \mathcal{F}$. Then,
(i) \hat{F} is necessarily a probability mass function $\{\hat{p}_1, \dots, \hat{p}_n\}$ on the observations Y_1, \dots, Y_n ;
(ii) \hat{g} is necessarily a natural cubic spline with knots at the observations t_1, \dots, t_n ; and
(iii) the penalized loglikelihood at the maximum reduces to the form

$$l_\alpha(\hat{g}, \hat{p}) = \frac{1}{n} \sum_{i=1}^n \{\log \hat{p}_i - b(\hat{g}_i, \hat{p}) + g_i Y_i\} - \frac{1}{2} \alpha \hat{g}^T K \hat{g},$$

$$\text{with } b(\hat{g}_i, \hat{p}) = \log \sum_{j=1}^n \exp \{\hat{g}_i Y_j\} \hat{p}_j, \quad i = 1, \dots, n,$$

where $\hat{p} = (\hat{p}_1, \dots, \hat{p}_n)^T$, K is a known, symmetric, banded matrix depending only on the covariates, and $\hat{g} = (\hat{g}_1, \dots, \hat{g}_n)^T$ denotes a vector of knot values. Note that $\sum_{i=1}^n p_i = 1$, so that the maximization problem involves at most $2n - 1$ variables, with this number being smaller when there are ties in the dataset.

Proof. For fixed g , Vardi (1985) showed that the maximizer \hat{F} of the unpenalized loglikelihood function (4) is necessarily a probability mass function $\{\hat{p}_1, \dots, \hat{p}_n\}$ on the observed support Y_1, \dots, Y_n . Thus, for fixed g , the loglikelihood portion of l_α at the maximum must be of the empirical form $n^{-1} \sum_{i=1}^n \{\log \hat{p}_i - b(g(t_i), \hat{p}) + g_i Y_i\}$ with $b(g(t_i), \hat{p}) = \log \sum_{j=1}^n \exp \{g(t_i) Y_j\} \hat{p}_j$, the empirical analogue of (3). On the other hand, for fixed distribution F , Green & Silverman (1994) showed that the maximizer of l_α over g must be a natural cubic spline with knots at the observed covariates, t_1, \dots, t_n . Moreover, the penalty term $\int \{g''(t)\}^2 dt$ necessarily reduces to the quadratic expression $\hat{g}^T K \hat{g}$, where $\hat{g} = (\hat{g}_1, \dots, \hat{g}_n)^T$ denotes the vector of knot values at the observed covariates and K is a matrix given in Green & Silverman (1994, Section 2.1.2) that depends only on the covariates. Combining these two arguments gives the result. \square

To translate the estimated natural cubic spline $\hat{g}(t)$ into a mean curve, define the estimated cumulant generating function by $b(\hat{g}(t), \hat{p}) = \log \sum_{i=1}^n \exp \{\hat{g}(t) Y_i\} \hat{p}_i$. The doubly-nonparametric estimator of the mean $\mu(t) = E(Y|t)$ and error distribution $F_t(y) = \Pr(Y \leq y|t)$ at covariate value t can then be computed as

$$\hat{\mu}(t) = \sum_{i=1}^n Y_i \exp\{-b(\hat{g}(t), \hat{p}) + \hat{g}(t) Y_i\} \hat{p}_i \quad \text{and}$$

$$\hat{F}_t(y) = \sum_{i=1}^n 1(Y_i \leq y) \exp\{-b(\hat{g}(t), \hat{p}) + \hat{g}(t) Y_i\} \hat{p}_i,$$

respectively. The fitted model also induces a variance function for the data, given by

$$\hat{V}(t) = \hat{\text{Var}}(Y|t) = \sum_{i=1}^n (Y_i - \hat{\mu}(t))^2 \exp\{-b(\hat{g}(t), \hat{p}) + \hat{g}(t) Y_i\} \hat{p}_i.$$

Variance functions are sometimes of direct interest (e.g., [Ruppert, Wand, Holst & Hössjer, 1997](#)), but usually play more of a secondary role in the computation of variability bands for inferences on the mean.

In all of the above, and throughout the rest of this paper, we treat the smoothing parameter α as being given. There are a few competing ways to choose the smoothing parameter in practice, with perhaps the most popular approach being cross-validation. We postpone discussion on practical selection of the smoothing parameter until Section 8.

It is worth mentioning here that the exponential tilt representation is also used in [Huang \(2014\)](#) to develop a semiparametric extension of GLMs in which the mean function is parametric but the response distribution is nonparametric. The current proposal is therefore an extension of this that allows both components to be nonparametric. Having a parametric form for the mean curve is particularly useful in biostatistical, agricultural and engineering contexts where explicit treatment effects are of interest, but the doubly-nonparametric approach can still be useful for model diagnosis and model selection in those scenarios. The proposed approach also extends the semiparametric proportional likelihood ratio model of [Luo & Tsai \(2012\)](#), and subsequent work by [Chan \(2013\)](#) and [Davidov & Ilipoulos \(2013\)](#), by letting the canonical parameter θ be an arbitrary smooth function of t rather than setting $\theta = t\beta$ for some β .

3 Implementation via existing algorithms

An attractive feature of the doubly-nonparametric approach is that it can be readily fitted using existing algorithms. More precisely, computations can be carried out via the following two steps. First, for fixed empirical distribution \mathbf{p} , optimization over \mathbf{g} can be achieved by Fisher scoring (see [Green & Silverman, 1994](#), Theorem 5.1), whereby a current estimate $\mathbf{g}^{(m)}$ of \mathbf{g} is updated via

$$\mathbf{g}^{(m+1)} = \left(W^{(m)} + \alpha K\right)^{-1} W^{(m)} \mathbf{g}^{(m)} + \left(W^{(m)} + \alpha K\right)^{-1} \left(\mathbf{Y} - \boldsymbol{\mu}^{(m)}\right). \quad (6)$$

Here, $\boldsymbol{\mu}^{(m)}$ is the current mean vector with components

$$\mu_i^{(m)} = \frac{\sum_{j=1}^n Y_j \exp\{g_i^{(m)} Y_j\} p_j}{\sum_{j=1}^n \exp\{g_i^{(m)} Y_j\} p_j},$$

and $W^{(m)}$ is a diagonal matrix of the current variances with

$$W_{ii}^{(m)} = \frac{\sum_{j=1}^n \left(Y_j - \mu_i^{(m)}\right)^2 \exp\{g_i^{(m)} Y_j\} p_j}{\sum_{j=1}^n \exp\{g_i^{(m)} Y_j\} p_j}.$$

Second, for fixed vector \mathbf{g} , it can be shown via Lagrangian multipliers (e.g. [Fokianos *et al.*, 2001](#)) that the optimizer in \mathbf{p} is of the form

$$p_i = \frac{1}{\sum_{k=1}^n \exp\{-b_k + g_k Y_i\}}, \quad (7)$$

where $\mathbf{b} = (b_1, \dots, b_n)^T$ satisfies

$$\sum_{j=1}^n \frac{\exp\{-b_i + g_i Y_j\}}{\sum_{k=1}^n \exp\{-b_k + g_k Y_j\}} = 1. \quad (8)$$

The doubly-nonparametric MLE can then be obtained by iterating between the two steps, (6) and (7)–(8); see [Green & Silverman \(1994\)](#) and [Fokianos *et al.* \(2001\)](#) for more details on the computation of the individual steps. For smaller sized problems, such as those with less than 100 observations, we found that maximizing over \mathbf{g} and \mathbf{p} jointly can work just as well as this iterative scheme, although for larger sample sizes the iterative scheme is more efficient.

4 Variability bands for the mean

For given F , it can be shown (c.f. [Hastie & Tibshirani, 1990](#), equation (6.26)) that the covariance matrix V_g of \hat{g} is approximately $V_g \approx [V + \alpha K]^{-1} V [V + \alpha K]^{-1}$, where $V = \text{Diag}(\text{Var}(Y_1), \dots, \text{Var}(Y_n))$ is a diagonal matrix of the variances. Moreover, each $\hat{g}_i - E(\hat{g}_i)$ is approximately $N(0, V_g^i)$ in distribution for large n , where V_g^i denotes the i th diagonal entry of V_g . Thus, approximate 95% pointwise upper and lower variability bounds for each $\hat{g}_i \equiv \hat{g}(t_i)$ can be computed as $g_i^U = \hat{g}_i + 2(V_g^i)^{1/2}$ and $g_i^L = \hat{g}_i - 2(V_g^i)^{1/2}$, respectively, as in classical smoothing spline GLMs. We can then translate these bounds into upper and lower pointwise variability bounds for each $\hat{\mu}_i$ via

$$\mu_i^U = \int y \exp \{ -b(g_i^U, F) + g_i^U y \} dF(y) \quad \text{and} \quad \mu_i^L = \int y \exp \{ -b(g_i^L, F) + g_i^L y \} dF(y),$$

respectively. For moderately large datasets, interpolating each pointwise upper or lower bound values using piecewise linear functions generally suffices to produce a decent picture of the variability around the fitted mean curve. Other more sophisticated methods exist, but it is not apparent that the associated accuracy gain are worth the additional computational costs.

Since F is not known in the doubly-nonparametric case, we perform the above calculations using the plug-in estimator \hat{F} instead to obtain approximate, pointwise, conditional 95% variability bands for the estimated mean. This is how the variability band in the LIDAR example in [Figure 1](#) is computed. The plug-in method seems to work reasonably well even for moderately small sample sizes, as our examples in [Section 6](#) indicate. This is consistent with the finding in [Huang & Rathouz \(2013\)](#) that the mean and error distribution in any GLM are always orthogonal. Note that variability bands can be interpreted as confidence bands if the target function is taken to be $E(\hat{g})$ rather than the true value g^* (see [Ruppert, Wand & Carroll, 2003](#), Section 6)

5 Two useful properties

5.1 Automatic preservation of support bounds

Suppose the support of the responses is constrained in a subset of \mathbb{R} of the form $(-\infty, U]$, $[L, U]$ or $[L, \infty)$, where L and/or U are not necessarily known. An attractive property of the proposed approach is that it automatically produces estimated mean curves and variability bands that are contained within the same interval. This is due to the exponential tilt mechanism which preserves the underlying support and the fact that the estimated distribution \hat{F} only places positive mass on the observed support.

5.2 Invariance to shift and scaling

The exponential tilt approach to GLMs is also invariant to location-shifts and scaling of the response. This is generally not true of GLMs with a given exponential family or QL models with a given variance function.

For example, suppose that instead of observing data from the Gamma regression model $Y_i | t_i \sim \text{Gamma} \{ \text{mean} = \mu(t_i), \text{var} = \phi \mu(t_i)^2 \}$, we observe $\tilde{Y}_i = aY_i + b$ for some scalars a and b . Then, it is easy to see that the observed data (t_i, \tilde{Y}_i) no longer come from a Gamma model. Moreover, the observed data do not come from the same QL family as the original data. Modeling such data with a classical Gamma GLM would require an appropriate recentering and rescaling of the data, a task that is not trivial if a and b are unknown. In contrast, the transformed data can be directly analyzed by the proposed approach, with the estimated

mean-curve and variability bands being precisely the appropriate shifted and scaled version of the mean-curve and variability bands for the original, unobserved data. The same is also true for shifting and scaling of the predictors t_i . The doubly-nonparametric approach is invariant to shifting and scaling of the data.

6 Estimation accuracy, coverage rates and robustness to model misspecification

The doubly-nonparametric GLM is an extension of the parametric GLMs of McCullagh & Nelder (1989), the nonparametric GLMs of Green & Silverman (1994), the semiparametric GLMs of Huang (2014) and the semiparametric proportional likelihood ratio models of Luo & Tsai (2012). Thus, the approach is expected to be flexible enough to handle a very wide range of scenarios. Here, we examine the practical performance of the proposed approach using various simulations. Recently, Huang & Rathouz (2013) showed that estimation and inferences for a parametric GLM can remain asymptotically efficient even in the presence of an infinite-dimensional error distribution parameter. Our simulation results here suggest that nonparametric estimation of the error distribution also has minimal effect on the estimation of and inferences on a nonparametric mean-curve.

In all our simulations, the covariates t_1, \dots, t_n are equally-spaced between 0 and 1 with a sample size $n = 80$, and the mean-curve is taken to be the “bump function” from Ruppert, Wand & Carroll (2003, Section 5.6.3), which is given by

$$E(Y|t) = \mu(t) = \frac{1}{0.1 + t} + 8 \exp \{ -400(x - 0.5)^2 \}. \quad (9)$$

The mean-curve (9) comprises of a sudden peak in the middle of an otherwise smoothly decaying trend and is a good example for testing the local and global properties of nonparametric regression methods. It is plotted as the dashed curve in Figure 2.

To examine practical performance in both correctly-specified and misspecified scenarios, we simulate data under two classical nonparametric GLM settings (settings 1 and 2) and two misspecified settings (settings 3 and 4):

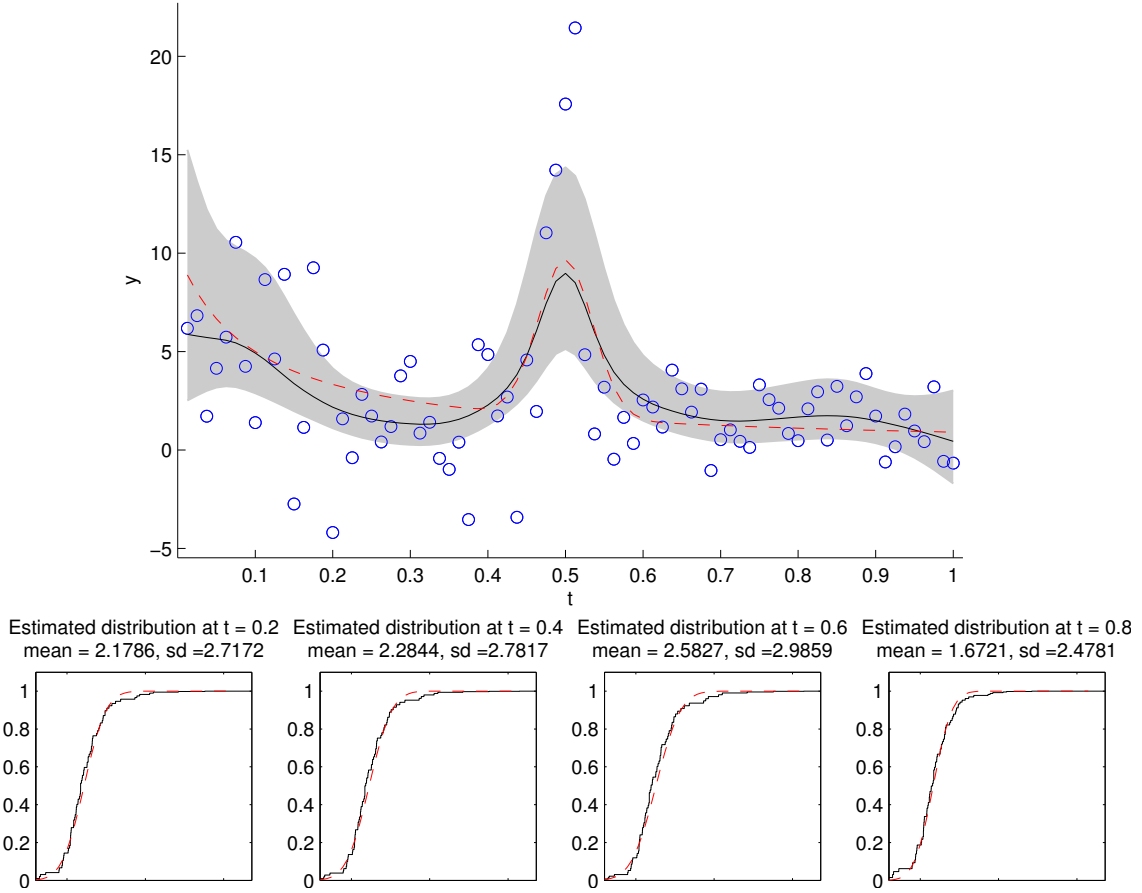
1. $Y|t \sim N\{\mu(t), 1\}$
2. $Y|t \sim \text{Poisson}\{\mu(t)\}$
3. $Y|t \sim \text{negative-binomial}\{\text{mean} = \mu(t), \text{variance} = \mu(t) + \mu(t)^2/\nu\}$ with $\nu = 10$
4. $Y|t \sim N\{\mu(t), 1 + 0.7\mu(t) + 0.5\mu^2(t)\}$

Setting 3 here reflects the popular use of the negative-binomial distribution as a model for overdispersed or underdispersed count data relative to the Poisson distribution. Note that the negative-binomial is not a GLM unless the dispersion parameter ν is known *a priori*, so that this setting is outside our model space. Note also that setting $\nu = 10$ here corresponds to moderate overdispersion. In Setting 4, the data come from a normal distribution but the variance is a quadratic function of the mean. This heteroscedastic normal family is also outside the model space for both classical nonparametric and the proposed doubly-nonparametric frameworks.

Each simulated dataset was modeled using (i) a nonparametric GLM assuming a constant variance function $V(\mu) = \sigma^2$, (ii) a nonparametric GLM assuming quasipoisson variance $V(\mu) = \phi\mu$, and (iii) a doubly-nonparametric GLM with no assumption on the variance function. Note that method (i) is correctly specified for Setting 1, method (ii) is correctly specified for Setting 2, and all three methods are misspecified for Settings 3 and 4.

The value of the smoothing parameter used in each approach was fixed at the value that minimized the average root (weighted) mean-square error (RMSE) across the simulations.

Figure 2: Fitted model (solid) and 95% variability band (shaded) for a heteroscedastic dataset from Setting 4, with true mean-curve (dashed) and response distributions (dashed) at the quintiles. Sample size $n = 80$, smoothing parameter $\alpha = 7.35 \times 10^{-6}$.



Recall that the root weighted mean-square error is defined by

$$\text{RMSE} = \sqrt{\frac{1}{n} \sum_{i=1}^n \frac{(\hat{\mu}_i - \mu_i^*)^2}{V_i^*}},$$

where for each observation i , $\hat{\mu}_i$ and μ_i^* are the fitted and true values of the mean, respectively, and V_i^* is its true variance. Of course, we never know the true mean-curve μ^* in practice, nor do we know the true variance V^* for assessing the appropriate goodness-of-fit, but the choice of smoothing parameter here allows us to make fair comparisons using the best possible fits under all three methods. The MATLAB code used to fit each model can be obtained by emailing the author.

Pointwise 95% variability bands for the mean-curve were then constructed using either the method in [Green & Silverman \(1994, Section 5\)](#) for classical nonparametric GLMs or the prescription in [Section 4](#) for the doubly-nonparametric approach. Coverage rates of these variability bands for the true mean at three locations, namely $t = 0.4, 0.5$ and 0.75 , corresponding to a trough, a sudden peak and an area of steady decline, respectively, are given in [Table 1](#). The table also displays the best RMSE achieved by each method.

We see from [Table 1](#) that the doubly-nonparametric approach can perform as well as a correctly-specified model for both estimation and inference. For misspecified models, its performance is consistently better than classical approaches with the wrong response distribution or variance function. In particular, coverage rates for classical nonparametric GLMs can be quite disastrous under model misspecification.

To gain some insight into how the doubly-nonparametric approach deals with model misspecification, the default graphical output from the MATLAB routine `dnpglm.m` for a

generic dataset under Setting 4 is given in Figure 2. We see that the doubly-nonparametric approach tries to find an exponential family that best approximates the $N(\mu, 1 + 0.7\mu^2 + 0.5\mu^2)$ subfamily across the range of means μ in the dataset. Of course, it can never replicate the true response distribution exactly because it is outside the model space, but it does a surprisingly good job in approximating it. This reinforces the versatility and flexibility of exponential families for modeling data, as suggested in Hiejima (1997).

7 Extension to multiple covariates

Two popular ways of extending classical nonparametric GLMs to multiple covariates are generalized additive models (Hastie & Tibshirani, 1990) and reducing the dimensionality of the problem using linear predictors (e.g., Green & Silverman, 1994, Section 6.4). These two approaches are directly applicable to the doubly-nonparametric case also, with minimal additional effort. For notational convenience, write $\mathbf{t}_i = (t_i^1, \dots, t_i^d)^T$ to be a d -vector of covariates for observation i .

7.1 Additive models

Additive models make the assumption that the natural parameter θ_i has the form $\theta_i = \sum_{j=1}^d g_j(t_i^j)$ for a set of smooth but otherwise unspecified functions $\{g_1, \dots, g_d\}$. That is, the effect of each covariate on the response is additive on the natural, canonical scale. The corresponding doubly-nonparametric penalized loglikelihood function for $\{g_j\}$ and F is then given by

$$l_{\alpha}(\{g_j\}, F) = \frac{1}{n} \sum_{i=1}^n \left\{ \log dF(Y_i) - b(\{g_j(\mathbf{t}_i)\}, F) + \sum_{j=1}^d g_j(t_i^j) Y_i \right\} - \frac{1}{2} \sum_{j=1}^d \alpha_j \int \{g_j''(t)\}^2 dt,$$

where

$$b(\{g_j(\mathbf{t}_i)\}, F) = \log \int \exp \left\{ \sum_{j=1}^d g_j(t_i^j) y \right\} dF(y), \quad i = 1 \dots, n.$$

Note that $\alpha = (\alpha_1, \dots, \alpha_d)^T$ is now a d -vector of smoothing parameters, one for each covariate.

By a similar argument to Proposition 1, maximizing this penalized loglikelihood over the $d + 1$ infinite-dimensional parameters in $\{g_j\}$ and F reduces to a finite optimization over at most $(d + 1)n - 1$ variables. As with classical nonparametric GLMs, additive models can quickly become computationally unfeasible as the number d of covariates increases. However, if some of the covariates only take on a small number of values, such as those that are binary or fixed by design, then the additional computational burden corresponding to those covariates is relatively minimal. For a general discussion on additive models, see Hastie & Tibshirani (1990).

7.2 Linear predictors and nonparametric links

A computationally less demanding approach to handle multiple covariates is to reduce dimensionality through the use of a linear predictor. For any vector $\gamma \in \mathbb{R}^d$, define $s_i(\gamma) = \mathbf{t}_i^T \gamma$ to be the associated linear combination of the covariates \mathbf{t}_i , which is called a linear predictor. The natural parameter θ_i of the exponential family is then assumed to be related to the linear predictor through $\theta_i = g(s_i(\gamma))$, where g is some smooth but unspecified function. Note that for each fixed γ , the problem reduces to the univariate problem of Section 2.2. Thus, the extension here corresponds an additional level of optimization in d dimensions over γ . This approach is essentially equivalent to using a nonparametric link function.

8 Choosing the smoothing parameter

8.1 Cross-validation via Pearson residuals

An excellent summary of the competing approaches to smoothing parameter selection in smoothing spline GLMs is given in [Green & Silverman \(1994, Section 3.1\)](#). The points raised there apply equally to the doubly-nonparametric approach. In particular, we are sympathetic to the philosophical view that the free choice of smoothing parameter is an advantageous feature of the method that allows different features of the data to be explored on different scales, and that “it may well be that such a subjective approach is in reality the most useful one”. However, a more pragmatic statistician may insist on the need for an automatic method, whereby the smoothing parameter is chosen by the data. To this end, cross-validation, perhaps the most popular method for automatic selection, is also applicable in the doubly-nonparametric setting. We discuss one approach to cross-validation here. We also make a brief comment on the sensitivity of the approach to the choice of smoothing parameter.

A preliminary step that we have found useful in both reducing computation time and to “standardize” the magnitude of the smoothing parameter across different problems is to first rescale both the covariates and responses to the unit interval $[0, 1]$. This rescaling, and back-transforming on to the original scale, is done internally by the MATLAB routine that fits the doubly-nonparametric GLM.

One way to implement cross-validation for selecting the smoothing parameter in the doubly-nonparametric framework is an adaptation of least-squares cross-validation for smoothing splines (e.g. [Green & Silverman, 1994, Section 3.2](#)) using Pearson residuals. Pearson residuals are more appropriate than ordinary residuals when the variance is not assumed constant. For given α , let $\hat{\mu}^{(-i)}(t; \alpha)$ and $\hat{V}^{(-i)}(t; \alpha)$ be the estimated mean curve and variance function, respectively, obtained by omitting observation i . A cross-validation score for α based on Pearson residuals can then be defined as

$$CV(\alpha) = \sum_{i=1}^n \frac{\{Y_i - \hat{\mu}^{(-i)}(t_i; \alpha)\}^2}{\hat{V}^{(-i)}(t_i; \alpha)}.$$

This cross-validation score can then be minimized over $\alpha > 0$, leading a data-dependent “optimal” choice for α . In practice, a grid search over a suitable range of α values often suffices.

As with any smoothing method, computational burden can be reduced by cross-validating more than one observation at a time. For example, two-fold cross-validation, in which half the observations are omitted at a time, would require only two passes through the data per split. The smoothing parameter used to produce [Figure 1](#) was chosen in this manner via minimizing a two-fold cross-validation score over α values between $0.01n^{-4/5}$ and $0.50n^{-4/5}$ in steps of $0.01n^{-4/5}$, where we have used the optimal rate $n^{-4/5}$ from penalized smoothing splines ([Claeskens, Krivobokova & Opsomer, 2009](#)) as a guide for the range of α to search over.

8.2 Sensitivity to the smoothing parameter

The smoothing parameter in doubly-nonparametric GLMs plays a similar role in controlling the smoothness of the mean curve as it does in classical smoothing spline nonparametric GLMs. Interestingly, we have found that its effect on the estimated response distributions to be much less pronounced. This perhaps alludes to a doubly-nonparametric analogue of the recent discovery in [Huang & Rathouz \(2013\)](#) which showed that a parametric mean function is orthogonal to the response distribution in any GLM. This also reinforces that there may be

little to lose by adopting the more general doubly-nonparametric framework over classical nonparametric GLMs.

9 Discussion

The smoothing spline estimator used in this paper is one of many approaches that can be used in the doubly-nonparametric framework for estimating the mean curve. An alternative method is to use low-ranked penalized splines with a fixed number of knots. This may reduce computational burden at the cost of decreased flexibility. Kernel-type smoothers, however, are not compatible with the doubly-nonparametric framework because they are usually explicitly constructed from data rather than via optimizing a (penalized) likelihood function. For the error distribution, empirical likelihood is also one of many approaches that can be used in the doubly-nonparametric framework. Other approaches include mixture models and series expansions. Again, finite (or truncated) versions of these methods can increase computational efficiency at the cost of decreased flexibility. The development of software to incorporate these variations is currently work in progress.

Acknowledgements

The author thanks Professor Peter Green for discussions that led to the writing of this paper and Professors Matt Wand, Paul J. Rathouz and Joe Guinness for comments that improved the paper.

Bibliography

- Chan, K. C. G. (2013) Nuisance parameter elimination for proportional likelihood ratio models with nonignorable missingness and random truncation. *Biometrika*, **100**, 269–276.
- Chiou, J. and Müller, H. (1998) Quasi-likelihood regression with unknown link and variance functions. *Journal of the American Statistical Association*, **93**, pp. 1376–1387.
- Claeskens, G., Krivobokova, T. and Opsomer, J. D. (2009) Asymptotic properties of penalized spline estimators. *Biometrika*, **96**, pp. 529–544.
- Davidov, O and Iliopoulos, G. (2013) Convergence of Luo and Tsai’s iterative algorithm for estimation in proportional likelihood ratio models. *Biometrika*, **100**, 778–780.
- Eicker, F. (1967) Limit theorems for regressions with unequal and dependent errors. *Proc. Fifth Berkeley Symp. on Math. Statist. and Prob.*, **1**, 59–82.
- Fokianos, K, Kedem, B, Qin, J. and Short, D. A. (2001) A semiparametric approach to the one-way layout. *Technometrics*, **43**, 56–65.
- Green, P. J. and Silverman, B. W. (1994) *Nonparametric Regression and Generalized Linear Models*. Boca Raton: Chapman and Hall.
- Hastie, T. J. and Tibshirani, R. J. (1990) *Generalized Additive Models*. Boca Raton: Chapman and Hall.
- Hiejima Y. (1997) Interpretation of the quasi-likelihood via the tilted exponential family. *Journal of the Japan Statistical Society*, **2**, 157–164.
- Huang, A. (2014) Joint estimation of the mean and error distribution in generalized linear models. *Journal of the American Statistical Association*, **109**, 186–196.
- Huang, A. and Rathouz, P. J. (2013) Orthogonality of the mean and error distribution in generalized linear models. *Communications in Statistics: Theory and Methods*. (to appear).
- Luo, X. and Tsai, W. Y. (2012). A proportional likelihood ratio model. *Biometrika*, **99**, 211–222.
- McCullagh, P. and Nelder, J. A (1989) *Generalized Linear Models*. 2nd edition. London: Chapman and Hall.
- Rathouz, P. J. and Gao, L. (2009) Generalized linear models with unspecified reference distribution. *Biostatistics*, **10**, pp. 205–218.
- Ruppert, D., Wand, M. P., and Carroll, R. J. (2003) *Semiparametric Regression*. New York: Cambridge University Press.

Table 1: Coverage rates (%) for pointwise 95% variability bands for the mean at $t = 0.4, 0.5$ and 0.75 , with corresponding average root weighted mean-squared errors (RMSE), for classical nonparametric (NP) and doubly-nonparametric (DNP) GLMs. $N=5,000$ simulations for each setting. Sample size $n = 80$ for each simulation.

method	working	1. Normal				2. Poisson			
	variance	$t=0.4$	$t=0.5$	$t=0.75$	RMSE	$t=0.4$	$t=0.5$	$t=0.75$	RMSE
NP	σ^2	96.8	53.8	96.2	0.475	86.2	27.9	97.5	0.417
	$\phi\mu$	99.2	99.1	96.7	0.414	88.1	87.1	95.8	0.371
DNP	—	92.6	52.1	93.9	0.456	89.5	79.2	93.5	0.382

method	working	3. Negative-binomial				4. Heteroscedastic normal			
	variance	$t=0.4$	$t=0.5$	$t=0.75$	RMSE	$t=0.4$	$t=0.5$	$t=0.75$	RMSE
NP	σ^2	82.9	26.4	98.8	0.398	64.4	23.1	97.2	0.383
	$\phi\mu$	75.3	69.7	90.8	0.356	30.5	30.2	53.4	0.341
DNP	—	83.8	80.2	94.5	0.354	71.5	68.4	96.7	0.346

Ruppert, D., Wand, M. P., Holst, U., & Hössjer, O. (1997) Local polynomial variance-function estimation. *Technometrics*. **39**, pp. 262–273.

Vardi, Y. (1985) Empirical distributions in selection bias models. *Annals of Statistics* **13**, pp. 178-203.

White, H. (1982) Maximum likelihood estimation of misspecified models. *Econometrica* **50**, pp. 1-25.